Open Science Journal of Statistics and Application
2013; X(X): XX-XX
Published online MM DD 2013 (http://www.openscienceonline.com/journal/osjsa)

Loglinear modelling of cancer patients cases in Nigeria: An exploratory study approach

Odetunmibi, O. A.[1], Adejumo, A. O.[2, *], O. O. M. Sanni[2]

[1]Department of Computer and Information Sciences, Covenant University, Otta, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria.

Email address

oluwole.odetunmibi@covenantuniversity.edu.ng (Odetunmibi, O. A.), aodejumo@unilorin.edu.ng (Adejumo, A. O.)

To cite this article

Odetunmibi, O. A., Adejumo, A. O., O. O. M. Sanni. Loglinear Modelling of Cancer Patients Cases in Nigeria: An Exploratory Study Approach. *Open Science Journal of Statistics and Application*. Vol. 1, No. 1, 2012, pp. 1-7.

Abstract

The spread of cancer disease today worldwide is becoming rampant. Curbing the menace that the disease pose to the humanity has been a thing of concern and has put all hands on deck. Those that are health related workers and non-health related workers. The main objectives of this research work are to: test whether treatment Outcome (O) of cancer patients is dependent of Age (A) and Gender (G) from the two hospitals we have; check for the best model among various models that we have from the two locations; compare the result of the two hospitals in order to be able to conclude whether treatment outcome is the same from the two locations; and use the result of the two hospitals to determine what happen to cancer patients in South West region of Nigeria. We observed that Model (AO:GO which uses Age: Outcome of treatment and Gender: Outcome of treatment) has the minimum Akaike Information Criteria (AIC) value from the two hospitals and therefore is accepted to be the best model. We also observed that Age and Gender are individually independent of treatment outcome of cancer patients from the two hospitals. We can therefore conclude that the treatment outcomes from the two hospitals are the same and this implies that South-West region of Nigeria has the same treatment outcome for cancer patients.

Keywords

Cancer, Loglinear, Exploration, Likelihood Ratio, Treatment Outcome

1. Introduction

Loglinear models are special cases of generalized linear models (GLM, which include regression and anova models) to better treat case of dichotomous and categorical variables. Loglinear analysis deals with association of categorical or grouped data, looking at all levels of possible main and interaction effects, comparing the saturated model with reduced models, with the primary purpose being to find the most parsimonious model which can account for cell frequencies in a table, that is, log linear model analysis is a non dependent procedure for accounting for the distribution of cases in a cross tabulation of categorical variables. Log linear analysis is a type of multi -way frequency analysis (Agresti, 2002; Adejumo, 2005).

Let $Y = (Y_1, Y_2, Y_3, \ldots Y_D)$ be categorical variables. Then a rectangular ($N \times D$) data matrix consisting of N observations on Y can be re-arrange as a D-way contingency table with cells defined by joint levels of the variables. Let $n_{ij \ldots t}$ denote the frequency for a cell $Y = (i, j, \ldots, t)$ and $n = (n_{ij\ldots t})$. Suppose that Y has a multinomial distribution with an unknown parameter $\Theta = \{\Theta_{ij\ldots t}\}$, where $\Theta_{ij\ldots t} \geq 0$ and $\sum \Theta_{ij..t} = 1$.

The log-linear model is expressed in the form

$$\text{Log } \Theta = X\lambda$$

Where X is a $D \times r$ design matrix and λ is an $r \times 1$ parameter vector.

When Y has a Poisson distribution, the log-linear model is re-written by:

$$\text{Log } M = X\lambda$$

Where M = (mij…t = NΘij…t) is the vector of expected frequencies.

The data examined in this research work is focusing on a major disease, called cancer. When there is low oxygen in the cells and low alkalinity, the body's cells become abnormal and abnormal multiplication of white blood cells always denote toxicity, imbalance, poison and foreign invasion of that which is unnatural or that which should not be in the body. Excessive white blood cell count is almost always indicative of cancer (WHO, 2005).

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Cancer cell can spread to other parts of the body through the blood and lymph system.

Cancer is not just one disease but many diseases. There are more than 100 different types of cancer. Most cancers are named for the organ or type of cell in which they start.

Although, this research work is not generally based on the causes of cancer, but having an understanding of what cause it as it has been stated above will go a long way in dealing with its spread.

However, this research work is much more concerned about some factors that affect the outcome of treatment of cancer patients in Nigeria.

2. Methodology

The log-linear model which is used in analysis of contingency tables are a generalized linear models for counted data and the variety of associations and interaction terms in log-linear models can easily be describe by goodness of fit tests. The methodology of log-linear model for analysis of contingency tables is described in many book such as Bishop et.al (1975); Everitt (1977); Agresti (2002); Adejumo, 2005 and so on.

Consider an *I x J* contingency table. The log-linear model is represented by:

$$\log(M_{ij}) = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12} \tag{1}$$

For all i and j, under the constraints of the λ term to sum to zero over any subscript such as:

$$\sum_{i=1}^{I} \lambda_i^1 = 0,\ \sum_{j=1}^{J} \lambda_j^2 = 0\,,\ \sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = 0 \tag{2}$$

The log-linear model given above is called the saturated model or full model for the statistical dependency between Y_1 and Y2.

By analogy with analysis of variance models, we define the overall mean by:

$$\lambda_0 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log M_{ij} \tag{3}$$

The main effects of Y_1 and Y_2 by

$$\lambda_i^1 = \frac{1}{J} \sum_{j=1}^{J} \log M_{ij} - \lambda_0 \tag{4}$$

$$\lambda_i^1 = \frac{1}{I} \sum_{j=1}^{J} \log M_{ij} - \lambda_0 \tag{5}$$

And the two-factor effect between Y_1 and Y_2 by

$$\lambda_{ij}^{12} = \log M_{ij} = (\lambda_i^1 + \lambda_j^2) - \lambda_0 \tag{6}$$

Then the main and two-factor effects are determined by the odds and odds ratio, and can be written by:

$$\lambda_i^1 = \frac{1}{IJ} \sum_{i^1=1}^{I} \sum_{j=1}^{J} \log \frac{M_{ij}}{M_{i^1 j}}, \tag{7}$$

$$\lambda_i^2 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j^1=1}^{J} \log \frac{M_{ij}}{M_{ij^1}} \tag{8}$$

And

$$\lambda_{ij}^{12} = \frac{1}{IJ} \sum_{i^1=1}^{I} \sum_{j^1=1}^{J} \log \frac{M_{ij} M_{ij}}{M_{i^1 j} M_{ij^1}} \tag{9}$$

For the independence model that Y_1 is statistically independent of Y_2, the cell probability Mij can be factorized into the product of marginal probabilities M_{i+} and M_{+j}, that is,

$$\text{Mij} = \text{Mi+ M+j},$$

Where $M_{i+} = \sum_{j=1}^{J} M_{ij}$ and $M_{+j} = \sum_{i=1}^{I} M_{ij}$. Then the two-factor effect is:

$$\lambda_{ij}^{12} = \frac{1}{IJ} \sum_{i^1=1}^{I} \sum_{j^1=1}^{J} \log \frac{M_{i+} M_{+j} M_{i^1+} M_{+j^1}}{M_{i^1+} M_{+j} M_{i+} M_{+j^1}} = 0\,, \tag{10}$$

So that the log-linear model for the independence model is expressed by:

$$\log M_{ij} = \lambda_0 + \lambda_i^1 + \lambda_j^2\,, \text{ for all i and j} \tag{11}$$

For an I x J x K contingency table, the saturated log-linear model for the contingency table is:

$$\log M_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{ik}^{13} + \lambda_{jk}^{23} + \lambda_{ijk}^{123} \tag{12}$$

for all i, j and k.

The λ terms satisfy the constraints;

$$\sum_{i=1}^{I} \lambda_i^1 = \sum_{j=1}^{J} \lambda_j^2 = \sum_{k=1}^{K} \lambda_k^3 = 0, \tag{13}$$

$$\sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = \ldots = \sum_{k=1}^{K} \lambda_{jk}^{23} = 0, \tag{14}$$

$$\sum_{i=1}^{I}\lambda_{ijk}^{123}=\sum_{j=1}^{J}\lambda_{ijk}^{123}=\sum_{k=1}^{K}\lambda_{ijk}^{123}=0, \quad (15)$$

We define the λ terms as follows:
The overall mean is given by:

$$\lambda_0=\frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\log M_{ijk.} \quad (16)$$

The main effects of Y_1, Y_2, and Y_3 are:

$$\lambda_i^1=\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\log M_{ijk.}-\lambda_0 \quad (17)$$

$$\lambda_j^2=\frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\log M_{ijk.}-\lambda_0 \quad (18)$$

$$\lambda_k^2=\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\log M_{ijk.}-\lambda_0 \quad (19)$$

Each interaction effect is given by:

$$\lambda_{ij}^{12}=\frac{1}{K}\sum_{k=1}^{K}\log M_{ijk.}-(\lambda_i^1+\lambda_j^2)-\lambda_0 \quad (20)$$

$$\lambda_{ik}^{13}=\frac{1}{J}\sum_{j=1}^{J}\log M_{ijk.}-(\lambda_i^1+\lambda_k^3)-\lambda_0 \quad (21)$$

$$\lambda_{jk}^{23}=\frac{1}{I}\sum_{i=1}^{I}\log M_{ijk.}-(\lambda_j^2+\lambda_k^3)-\lambda_0 \quad (22)$$

and,

$$\lambda_{ijk}^{123}=\log M_{ijk.}-(\lambda_{ij}^{12}+\lambda_{ik}^{13}+\lambda_{jk}^{23})-(\lambda_i^1+\lambda_j^2+\lambda_k^3)-\lambda_0 \quad (23)$$

2.1. Model Selection in Categorical Data Analysis

Given a data set, several completing models may be ranked according to their Akaike Information Criteria (AIC) value with the one having the lowest AIC value being the best (See Galindo-Garre and Vermunt, 2005; Vermunt, 1997).

Suppose we denote two variables by a_0 and a_1 and suppose they have c_0 and c_1 categories respectively, we assumed that each variable a_j takes values $a_j = 1,.,.,., c_j$.

Let $P(a_0,a_1)$ be the probability that the variables a_0 and a_1 take values I_0 and I_1 respectively, and let $n(a_0,a_1$ be the corresponding cell frequency. If we denote the sample size by n then:

$$\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}P(a_0,a_1)=1$$
$$\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}n(a_0,a_1)=n \quad (24)$$

And the multinomial distribution probability mass function can be obtained by:

$$M(\{n(a_0,a_1)\}/\{P(a_0,a_1)\})=\frac{n!}{\prod_{a_0=1}^{c_0}\prod_{a_1=1}^{c_1}n(a_0,a_1)!}\prod_{a_0=1}^{c_0}\prod_{a_1=1}^{c_1}P(a_0,a_1)^{n(a_0.a_1)} \quad (25)$$

When equation above is regarded as a function of $P(a_0,a_1)$, then it is called likelihood function.

By ignoring $P(a_0,a_1)$, the log likelihood is given by:

$$l(\{P(a_0,a_1)\})=\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}n(a_0,a_1)\log P(a_0,a_1) \quad (26)$$

The alternative model having no restriction is represented as MODEL (0):

$$P(a_0,a_1)=b(a_0,.)b(.,a_1) \quad (27)$$

Where $b(a_0,.)=\sum_{a_0=1}^{c_0}P(a_0,a_1)$ and

$$b(.,a_1)=\sum_{a_1=1}^{c_1}P(a_0,a_1)$$

Substituting this, it is seen that this model has only one constraint:

$$\sum_{a_0=1}^{c_0}b(a_0,.)=\sum_{a_1=1}^{c_1}b(.,a_1)=1 \quad (28)$$

The alternative model having no restriction is represented as model (1):

$$P(a_0,a_1)=b(a_0,a_1)\,0 \quad (29)$$

Substituting this we have:

$$\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}b(a_0,a_1)=1 \quad (30)$$

It is also seen that this model has only one constraint.

We then find the maximum likelihood estimators (MLE's) for each model.

By substitution, the log likelihood for Model (0) is obtained as:

$$a_0(b(a_0,.)b(.,a_1)=\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}n(a_0,.)b(.,a_1)=1 \quad (31)$$

If we let $\log b(a_0,.)b(.,a_1)=A$

Then:

$$A=\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}n(a_0,a_1)\log b(a_0,.)+\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1}n(a_0,a_1)\log b(.,a_1)$$
$$A=\sum_{a_0=1}^{c_0}n(a_0,.)\log b(a_0,.)+\sum_{a_1=1}^{c_1}n(.,a_1)\log b(.,a_1) \quad (32)$$

Using langrange's multiple $\lambda_1,\lambda_2,\ldots$ we have

$$A=\sum_{a_0=1}^{c_0} n(a_0,.)\log b(a_0,.)+\sum_{a1=1}^{c_1} n(.,a_1)\log b(.,a_1)-\lambda_1\{\sum_{a_0=1}^{c_0} b(a_0,.)-1\}-\lambda_2\{\sum_{a_1=1}^{c_1} b(.,a_1)$$

Differentiating with respect to $b(a_0,.)$ and $b(.,a_1)$, λ_1 and λ_2 we get:

$$\frac{\delta A}{\delta b(a_0,.)}=\frac{n(a_0,.)}{b(a_0,.)}-\lambda_1 \tag{33}$$

$$\frac{\delta A}{\delta b(.,a_1)}=\frac{n(.,a_1)}{b(.,a_1)}-\lambda_2 \tag{34}$$

$$\frac{\delta A}{\delta \lambda_1}=-\left[\sum_{a_0}^{c_0} b(a_0,.)-1\right] \tag{35}$$

$$\frac{\delta A}{\delta \lambda_2}=-\left[\sum_{a_1}^{c_1} b(.,a_1)-1\right] \tag{36}$$

Equating the derivatives to zero and solving, we have:

$$\begin{aligned} n(a_0,.)&=\lambda_1 b(a_0,.)\\ n(.,a_1)&=\lambda_2 b(.,a_1)\end{aligned} \tag{37}$$

$$\sum_{a_0=1}^{c_0} b(a_0,.)=1$$

$$\sum_{a_1=1}^{c_1} b(.,a_1)=1 \tag{38}$$

$$\sum_{a_0=1}^{c_0} b(a_0,.)=\sum_{a_0=1}^{c_0} b(a_0,.)\lambda_1 b(a_0,.)$$

$$\mathrm{n}=\lambda 1,\ \sum_{a_0=1}^{c_0} b(a_0,.)=\lambda 1 \tag{39}$$

Similarly,

$$\sum_{a_1=1}^{c_1} n(.,a_1)=\sum_{a_1=1}^{c_1}\lambda_2 b(.,a_1) \tag{40}$$

$$n=\lambda_2\sum_{a_1-1}^{c_1} b(.,a_1)=\lambda_2 \tag{41}$$

$$b(a_o,.)=\frac{n(a_o,.)}{n}$$

By substituting all this, we have the maximum log likelihood to be:

$$\ell(\ \bar{b}(a_o,.)\bar{b}(.,a_1)\ )=\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1} n(a_0,a_1)\frac{\log n(a_0,.)n(.,a_1)}{n^2} \tag{42}$$

Then, from the two constrain the number of free parameter is equal to $(c_0\text{-}1) + (c_1\text{-}1)$, and hence the corresponding AIC is given by:

$$AIC(0)=-2\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1} n(a_0,a_1)\frac{\log n(a_0,.)n(.,a_1)}{n^2}+2(c_0+c_1-2) \tag{43}$$

Also doing the same thing for MODEL (1) we have AIC to be:

$$AIC(1)=-2\sum_{a_0=1}^{c_0}\sum_{a_1=1}^{c_1} n(a_0,a_1)\frac{\log n(a_0,a_1)}{n}+2(c_0c_1-1) \tag{44}$$

The data under consideration, the main factors are Age (A) with five levels, Gender (G) has two levels, and Treatment Outcome (O) having two categories as well. The following loglinear models are considered (see Clogg and Eliason, 1987; and Christensen, 1997).

Model 1: When no association exist between the variables (Model A.G.O), we have

$$\log(m_{ijk})=\lambda+\lambda_i^A+\lambda_j^G+\lambda_k^O \tag{45}$$

Model 2: Two ways association between just two variables (Model AG.O), the model is

$$\log(m_{ijkl})=\lambda+\lambda_i^A+\lambda_j^G+\lambda_k^O+\lambda_{ij}^{AG} \tag{46}$$

Model 3: Two ways association between each variable with Outcome (Model AO.GO), the model is

$$\log(m_{ijkl})=\lambda+\lambda_i^A+\lambda_j^G+\lambda_k^O+\lambda_{ik}^{AO}+\lambda_{jk}^{GO} \tag{47}$$

Model 4: Two ways association between the variables (Model AG. AO.GO), the model is

$$\log(m_{ijkl})=\lambda+\lambda_i^A+\lambda_j^G+\lambda_k^O+\lambda_{ij}^{AG}+\lambda_{ik}^{AO}+\lambda_{jk}^{GO} \tag{48}$$

Model 5: Three way association between the variables (Model AGO), the model is also called Saturated model and is

$$\log(m_{ijk})=\lambda+\lambda_i^A+\lambda_j^G+\lambda_k^O+\lambda_{ij}^{AG}+\lambda_{ik}^{AO}+\lambda_{jk}^{GO}+\lambda_{ijk}^{AGO} \tag{49}$$

Models in Equations (45) to (49) can be fitted using any of the following methods for the Maximum likelihood (ML) estimation; Newton-Raphson, Fishers Scoring or Iterative proportional fitting (IPF). However, in this work, Fishers scoring iterative method will be used (Christensen, 1997; Adejumo, 2005; Adejumo and Arabi, 2013). The equation of the Fisher-scoring algorithm is

$$(X^1W^{(k)}X)\lambda^{(k+1)}=X^1W^{(k)}Z^{(k)}$$

This is the likelihood equation of a generalized linear model with the response vector $Z^{(k)}$ and a random error covariate matrix $(W^{(k)})^{-1}$. If rank(X)=p holds, we obtain the ML estimate λ as the limit of

$\lambda^{(k+1)} = (X^1 W^{(k)} X)^{-1} X^1 W^{(k)} Z^{(k)}$.

An algorithm was written in R programme to evaluate the process described and the parameters, log-likelihood ratio statistic (G^2) as well as the Akaike information criteria (AIC) are obtained for each loglinear models stated above.

2.2. Analysis, Results and Discussion

The data on cancer patients who received treatment over the period of four years and nine months from Eko Hospital Lagos and Obafemi Awolowo University Teaching Hospital Ile-Ife (OAUTH).

A total number of 565 patients were recorded from Eko Hospital Lagos while a total of 750 patients were recorded from OAUTH Ile-Ife.

Table 1. *likelihood ratio test for hierarchical log-linear models*

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev)
Model 4	13.587365	13			
Model 3	10.618090	9	2.969274	4	0.56298
Model 2	3.916764	8	6.701327	1	0.00963
Model 1	0.596732	4	3.320032	4	0.50576
Saturated	0.000000	0	0.596732	4	0.96343

Table 2. *Aic values for the models*

MODEL	G^2	AIC	P-VALUE	Df
Model 1	0.59673	122.05	0.9634261	4
Model 2	3.9168	117.37	0.8645541	8
Model 3	10.618	122.67	0.3027975	9
Model 4	13.587	127.04	0.4035316	13
Saturated	0	129.45	1	0

Table 3. *Lr test for hierarchical log-linear models*

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev)
Model 4	23.720853	13			
Model 3	19.273721	9	4.447132	4	0.34886
Model 2	8.626868	8	10.646853	1	0.00110
Model 1	3.644471	4	4.982397	4	0.28911
Saturated	0.000000	0	3.644471	4	0.45625

able 4. *Aic Values for the Models*

MODEL	G^2	AIC	P-VALUE	Df
MODEL 1	3.644471	136.42	0.4562538	4
MODEL 2	8.626868	129.05	0.3747434	8
MODEL 3	19.27372	137.70	0.16296352	9
MODEL 4	23.72085	134.14	0.07379699	13
Saturated	0	136.42	1	0

Table 2 gives the values of AIC for each model and compared the values of each.

Tables 1 and 3 show the hierarchical of models on treatment outcomes of cancer patients from Eko Hospital Lagos and Obafemi Awolowo University Teaching Hospital (OAUTH) Ile-Ife and it also compares the model with the degree of freedom. The differences in model 1 and model 2 measures the distance of the best fit of model 2 from the best fit of model 1. This also applies to other models compared as shown in the tables.

3. Conclusion

The results we have from the analysis carried out from the data collected from each hospitals, using both the AICs and Log likelihood Ratios, give us the same result. accepting model AO.GO to be the best model and having each variable to be independent of one another, we can therefore concluded that the treatment outcomes from the two hospitals are the same and also that south-west region of the country has the same treatment outcome for cancer patients.

References

[1] Adejumo, A. O (2005), *Modelling Generalized Linear (Loglinear) Model for Raters Agreement measures;* Published by Peter Lang, Frankfurt am Main. (http://www.peterlang.de).

[2] Adejumo, A. O. and Arabi, F. J. (2013). "A Study of HIV Sero-Prevalence Surveillance

[3] Survey Data Using Loglinear Model. *Global Journal of Pure and Applied Mathematics.* Volume 9, Number 1: 73-81 India.

[4] Agresti, A. (2002). Categorical Data Analysis. John Wiley and Sons. 2nd Edition. New York.

[5] Bishop, Y. M. M, Feinberg, S. E. Holland, P. W. (1975). Discrete Multivariate Analysis. Cambridge MA; MIT Press.

[6] Christensen, R. (1997). Loglinear Models and Logistics Regression. Springer-Verlag Inc. NeW York. USA.

[7] Clogg C. C. and Eliason S. R. (1987) Some common problems in log-linear analysis, Sociological. Methods & Research, 16, 8-44.

[8] Everitt, B. S. (1992). The Analysis of Contingency Tables, London: Chapman & Hall.

[9] Galindo-Garre, F., and Vermunt, J.K. (2005). Testing Log-Linear Models with Inequality Constraints: A Comparison of Asymptotic, Bootstrap, and Posterior Predictive P Values. *Statistica Neerlandica, 59*(1), 82-94

[10] Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. London: Sage.

[11] World Health Organization (WHO 2005). WHO Publication on Cancer Diseases *WHO report 2005*. Geneva.